A Convolutional Neural Network for Enhancing the Detection of SSVEP in the Presence of Competing Stimuli

Aravind Ravi, *Student Member, IEEE*, Jacob Manuel, Nargess Heydari and Ning Jiang, *Senior Member, IEEE**

***Abstract*— Stimulus proximity has been shown to have an influence on the classification performance of a steady-state visual evoked potential based brain-computer interface (SSVEP-BCI). Multiple visual stimuli placed close to each other compete for neural representations leading to the effect of competing stimuli. In this study, we propose a convolutional neural network (CNN) based classification method to enhance the detection accuracy of SSVEP in the presence of competing stimuli. A seven-class SSVEP dataset from ten healthy participants was used for evaluating the performance of the proposed method. The results were compared with the classic canonical correlation analysis (CCA) detection algorithm. We investigated whether the CNN parameters learned on one inter-stimulus distance (ISD) can generalize across to other ISDs and sessions. The proposed CNN obtained a significantly higher classification accuracy than CCA in both the offline (75.3% vs. 67.9%, $(p < 10^{-3})$) and the simulated online (71.3% vs. 60.7%, $(p < 10^{-3})$) conditions for the closest ISD. The results suggest the following: the CNN is robust in decoding SSVEP across different ISDs, and can be trained independent of the ISD resulting in a model that generalizes to other ISDs.**

I. INTRODUCTION

Brain Computer Interfaces (BCIs) allow users to interact with their external environment by modulating their neural activity. BCI systems have the potential to benefit people with severe disabilities and assist them in their day-to-day activities. Among electroencephalogram (EEG)-based BCIs, steady state visual evoked potentials (SSVEP) are widely used for their high signal-to-noise ratio and low participant training time. SSVEP responses are usually elicited in the occipitoparietal area of the cortex when a user focuses his/her visual attention on a flickering light source characterized by a specific flicker frequency. These responses appear as an increase in the amplitude at the flicker frequency and potentially its harmonics in the EEG signal from the occipitoparietal area. By recognizing these characterizing frequencies in the EEG signal, the target stimulus focused by the user can be inferred.

Various factors such as stimulus frequency, color, and stimulus proximity have been shown to have an influence on the overall performance of an SSVEP-BCI [1]. Multiple visual stimuli placed close to each other compete for neural representations leading to the effect of competing stimuli [2], where a positive correlation between change in inter-stimulus distance (ISD) and the overall accuracy was found. These results were based on power spectrum features extracted on channels O1, O2 and Oz with Support Vector Machine (SVM) classifier. It was suggested that further investigations could be performed to identify better classification algorithms to enhance the SSVEP decoding accuracy in the presence of significant competing stimuli.

Recent developments in the field of deep learning have given rise to an increased interest in applying these techniques for biomedical signal analysis. In particular, convolutional neural networks (CNNs) have been the most widely used classification technique [3]. Studies show that CNNs can significantly improve SSVEP classification when compared to traditional methods. In [4], SSVEP signals are initially converted to the Fast Fourier Transform (FFT) representation prior to applying them as input to a CNN as SSVEP responses manifest as increase in amplitude at specific frequencies. Another study provides the time domain signal directly as input to a CNN for SSVEP classification [5]. In comparison to the CNNs using time-domain inputs, a CNN using frequency-domain inputs would have a similar but relatively simple network structure, which means relatively reduced computational complexity and reduced number of tunable parameters.

The current study proposes a variant of the CNN architecture proposed in [4] to enhance the overall SSVEP decoding performance, and in particular on the robustness in the presence of competing stimuli. We compared the performance of the proposed algorithm with Canonical Correlation Analysis (CCA) which is the most widely used classification method for SSVEP. The paper is organized as follows: In Section II, the methodology such as stimulus design, experimental protocol, data analysis, and proposed algorithm are presented. The experimental results are detailed in Section III. Section IV discusses the results and concludes the paper.

II. METHODOLOGY

A. Stimulus Design

A total of seven flickering stimuli were presented on an LCD screen having a refresh rate of 60Hz. They were designed based on the methods proposed in [6] and [7]. The frequencies of the seven stimuli were: 8.423Hz, 9.375Hz,

*Research is supported by an NSERC-CREATE grant (509950) and an NSERC-ENGAGE grant (401261605)

Aravind Ravi, Jacob Manuel, Nargess Heydari and Ning Jiang are with the Engineering Bionics Lab, Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L3G1 Canada. Email: aravind.ravi@uwaterloo.ca, jacob.manuel@uwaterloo.ca, nheydaribeni@uwaterloo.ca, ning.jiang@uwaterloo.ca

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

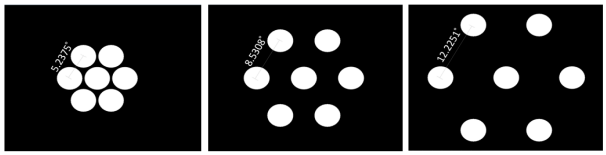

Fig. 1. Stimulus Configurations S1, S2 and S3

9.961Hz, 10.84Hz, 11.87Hz, 13.4Hz, and 14.87Hz, respectively. The stimuli were all circular in shape and white in color as this has been shown to elicit higher responses [8]. To evaluate the effects of changing ISDs, three different stimuli configurations (S1, S2 and S3) were designed, which are illustrated in Fig. 1. One stimulus locates in the center of the field-of-view, and six peripheral stimuli were placed concentrically around the central stimulus.The viewing angles of the peripheral stimuli were $5.24°$, $8.53°$ and $12.2°$ (0.6m between participant's eyes and the monitor) for S1, S2 and S3, respectively.

B. Data Acquisition and Experimental Protocol

The g.USBamp and Gammabox (g.tec Guger Technologies, Austria) wet electrode (g.Scarabeo) system were used to acquire EEG signals. The sampling rate was 1200Hz, and the electrodes were placed at O1, O2, Oz, PO3, POz, PO4 and FPz positions, according to the International 10-20 system. FPz was used as the ground and the reference electrode was placed on the right ear lobe.

EEG data was collected from ten healthy adults (eight male and two female, age 21-27) with normal or corrected-to-normal vision. Prior to the start of the experiment, each participant signed a Written Informed Consent. The experiment was approved by the Office of Research Ethics of the University of Waterloo (ORE #23152).

All participants were seated in a comfortable chair at 0.6m from the monitor. At the beginning of each trial, the participant was asked to gaze at a specific stimulus on the screen as directed by a yellow visual cue placed above the stimulus shown for 2s. Following the cuing period, a stimulation period followed for 6s, during which the participant would focus on the targeted stimulus on screen for the entire duration of the stimulation. A 4s break was provided before the next trial. One run, i.e. a continuous EEG recording, consisted of a total of 56 trials, as each of the seven flickering stimuli would be repeated for eight times in a randomized sequence. A total of three runs were performed, one for each stimulus configuration. Several minutes of resting period were provided between the runs. The order of the three runs was also randomized for each participant. They were asked to avoid eye blinks or any sudden jerky movements during each trial. The flickering stimuli and experimental protocol were implemented with OpenViBE. All data were recorded, stored and analyzed offline.

MATLAB (MathWorks Inc., Natick, USA) was used for all offline analysis. For pre-processing, the signals from channels O1, Oz and O2 were filtered by a 4th order Butterworth band-pass filter between 1Hz and 40Hz to remove any high frequency noise.

C. Classification based on Convolutional Neural Network

The data of each trial was segmented with a 1s-long sliding window and a step size of 100ms. Every segment consisted of 3 channels with 1200 samples per channel. Each segment was transformed into the magnitude spectrum representation by applying the FFT along each channel with a resolution of 0.2930Hz. The frequency components between 3Hz and 35Hz were considered as input data to the CNN as this band consists of relevant SSVEP information for this study. Therefore, the input to the CNN was a matrix of dimension $N_{ch} \times N_{fc}$ where N_{ch} was the number of EEG channels $(N_{ch} = 3)$ and N_{fc} was the number of frequency components $(N_{fc} = 110)$.

The proposed CNN has four main layers, an input layer, two convolutional layers, and a fully connected output layer. Each convolutional layer was composed of several feature maps that operate on the frequency information from different channels. The architecture proposed in this work was designed based on an intuitive understanding of EEG signal processing methods and the structure was a variant of the one proposed in [4]. Fig. 2 illustrates the proposed CNN architecture used in this study. $I(c, f)$ represents the input layer of the network with $1 \leq c \leq N_{ch}$ and $1 \leq f \leq N_{fc}$. The convolutional layers C1 and C2 comprised of $2 * N_{ch}$ feature maps each. The first convolutional layer, C1, had been designed to learn a linear combination of the EEG channels with the intuition of spatial filtering in EEG. This layer was designed to learn to weight the contribution of each channel differently. The kernel size for this layer was $N_{ch} \times 1$. Each feature map in C1 was of length N_{fc}. The C2 layer was designed to extract the features along the spectral representations of the input. The kernel had a size of 1×10 and each feature map in C2 was composed of 101 units. The output layer was a fully connected layer with a softmax function; it consisted of seven units, which represented the seven classes of SSVEP targets. Both C1 and C2 were followed by a batch normalization layer. Batch normalization and Dropout were added as they were shown to improve the performance, training speed and generalization of neural networks [9] [10]. The rectified linear unit (ReLU) activation function was chosen for all layers.

The stochastic gradient descent with momentum was used as the optimizer for the network with the categorical cross-entropy loss as the objective function. The weights of the network were initialized based on a Gaussian distribution $\mathcal{N}(0, 0.01)$. The hyper-parameters were chosen based on a grid search with the search

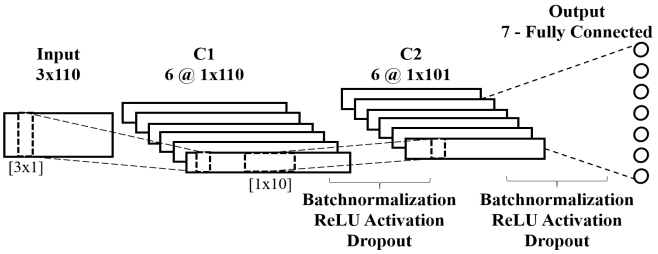

Fig. 2. Proposed Convolutional Neural Network Architecture

space defined as: Batch size (B): $2^b, b \in \{5, 6, 7, 8, 9\}$, Dropout ratio ($D$): $\{0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$, Number of Epochs ($E$): $\{20, 30, 40, 50, 60\}$, Learning Rate (α): $\{0.001, 0.002, 0.005, 0.01, 0.1\}$. The final parameters of the network were chosen as: $\alpha = 0.001$, $momentum = 0.9$, $D = 0.25$, $B = 256$ and $E = 40$, as it resulted in generally best performance for data from all participants.

D. Classification based on Canonical Correlation Analysis

CCA was performed on each segment of the pre-processed EEG data to identify the correlation between the EEG data and a reference signal. This technique was shown to produce superior performance in detecting SSVEP responses in EEG [11]. Consider two multidimensional variables X, Y where X refers to the set of multi-channel EEG data and Y refers to the set of reference signals of the same length as X. The linear combinations of X and Y are given as $x = X'W_x$ and $y = Y'W_y$. CCA finds the weights, W_x and W_y that maximize the correlation between x and y by solving (1). The maximum of ρ with respect to W_x and W_y is the maximum correlation. The reference signals Y_n were defined as (2), where $Y_n \in R^{2N_h \times N_s}$, f_n was the stimulation frequency, f_s was the sampling frequency, N_s was number of samples, and $N_h = 2$ was the number of harmonics used.

$$\max_{W_x, W_y} \rho(x, y) = \frac{E[W_x' X Y' W_y]}{\sqrt{E[W_x' X X' W_x] E[W_y' Y Y' W_y]}} \quad (1)$$

$$Y_n = \begin{bmatrix} \sin(2\pi f_n t) \\ \cos(2\pi f_n t) \\ \sin(4\pi f_n t) \\ \cos(4\pi f_n t) \end{bmatrix}, t = \begin{bmatrix} \frac{1}{f_s} \frac{2}{f_s} \dots \frac{N_s}{f_s} \end{bmatrix} \quad (2)$$

The canonical correlation features ρ_{f_i}, where $i = 1, 2, ..., 7$ were extracted for each segment of the EEG data, and the output class C for a given sample was determined as: $C = \text{argmax}(\rho_{f_i})$

E. Performance Evaluation

The classification performance of the proposed CNN architecture was compared against the state-of-the-art CCA method. Two types of analyses were performed: offline analysis and simulated online analysis. Offline analysis of the CNN was done to investigate whether parameters learned on one stimulus configuration can generalize across other stimuli configurations and sessions. Therefore, three cases were assessed. Case 1: CNN trained on S1 and tested on S2 and S3; Case 2: CNN trained on S2 and tested on S1 and S3; and Case 3: CNN trained on S3 and tested on S1 and S2. For all cases, a user-dependent training approach was used. As a result, for each stimulus configuration, two test-set accuracies were computed, and the mean accuracy was calculated for each participant. For example, test-set accuracy for S2 was calculated from Case 1 and Case 3 and the mean of the two test accuracies were calculated. These results were compared with the accuracies calculated using the classical-CCA method.

In addition to the offline analysis, a simulated online analysis was carried out to measure the real-time performance of the CNN. All three cases as outlined previously were tested. For the test data, the initial 1s of each trial [0.5s 1.5s] from the start of the flickering period was considered, whereas the training data was segmented in the same manner as mentioned earlier. The classification accuracies and the information transfer rate (ITR) were calculated for the CCA and the proposed CNN methods individually.

F. Statistical Analysis

Statistical analysis was performed to compare the performance of the two classification algorithms. From the eight trials per stimulus frequency, a leave-one-trial-out based partition was performed for each class, resulting in eight partitions. Each partition contained seven trials of EEG data per stimulus frequency. The test accuracy was calculated for each partition for all participants.

The overall performance of the two classification algorithms was analyzed using a mixed-effect model ANOVA. The response variable was the classification accuracy. The participant was a random factor, the stimulus configuration was a fixed factor with three levels (S1, S2, S3); the classification algorithm was a fixed factor with two levels (CNN and CCA). The null hypothesis was that the average classification accuracies were same for both algorithms. The secondary hypothesis was that the stimulus distance has no effect on the performance. The significance level was set as 0.05. These tests were common for both offline and simulated online analysis.

III. Results

Fig. 3 illustrates an example of the SSVEP responses of four consecutive stimuli on channel Oz for S1. The stimulus frequencies can be readily identified from the spectrogram. Fig. 4A presents the average classification accuracy across all participants for each stimulus configuration S1, S2 and S3, and shows the comparison between CNN and CCA. It is evident that the CNN outperformed CCA for all stimulus configurations with the largest improvement of over 7% for S1. The mixed-effect model ANOVA revealed a significant difference between the two algorithms ($p < 10^{-3}$), as well as a significant interaction between the algorithm and stimulus configuration ($p = 0.012$). Bonferroni simultaneous tests indicated that the CNN obtained a significantly higher

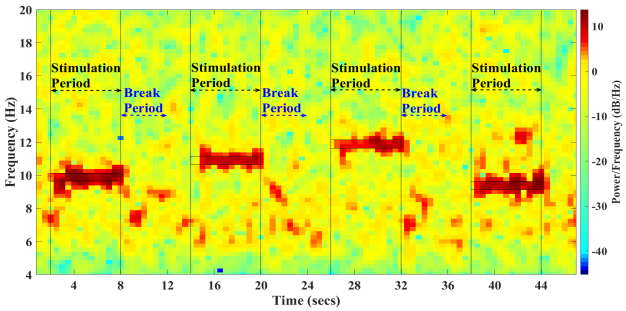

Fig. 3. SSVEP responses of four consecutive trials at frequencies 9.961Hz, 10.84Hz, 11.87Hz, 9.375Hz on channel Oz for S1.

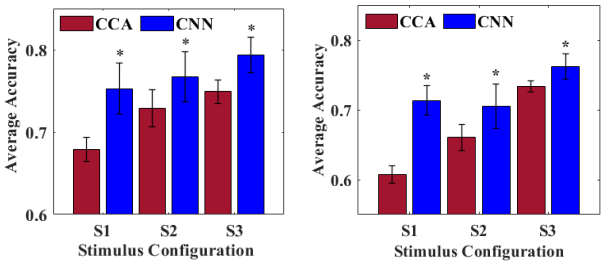

Fig. 4. A. Offline (Left); B. Simulated Online (Right); Average accuracies across all participants for each stimulus configuration ($*p < 0.05$).

accuracy than CCA-based classification for all stimulus distances ($p < 10^{-3}$), and accuracy of the three stimulus configurations were significantly different: $S3 > S2 > S1$, (all $p < 10^{-3}$).

Similar analysis was performed to evaluate the performance of the proposed approach in a simulated online setting. Fig. 4B illustrates the average accuracy for both algorithms across all participants. The proposed CNN achieved higher accuracy (S1: 71.3% vs. 60.7%, S2: 70.5% vs. 66.0%, S3: 76.2% vs. 73.4%) and ITR (S1: 51.0 bits/min vs. 34.4 bits/min, S2: 51.3 bits/min vs. 42.6 bits/min, S3: 59.0 bits/min vs. 52.5 bits/min) than CCA. The advantage of CNN over CCA is most pronounced in S1, where CNN has 10.6% accuracy gain and 16.6 bits/min in ITR than CCA. The statistical tests for the simulated online analysis showed CNN significantly outperformed CCA in accuracy ($p < 10^{-3}$), and obtained a significantly higher ITR ($p = 0.003$).

We also performed analysis on the computational load of the proposed CNN. The model was trained on an Intel Core i5-8400 CPU @ 2.80 GHz and 8 GB RAM. MATLAB's Deep Learning Toolbox was used to implement the model. The total number of trainable parameters were 4663. The overall training time was 6s. The inference time was measured as the time taken for the learned model to predict the class of a single segment of the test data. The mean inference time for all segments was found to be 1.3ms.

IV. Discussions and Conclusion

This study proposed a CNN-based classification method to enhance the decoding performance of a SSVEP-based BCI in the presence of competing stimuli with variable ISDs. The performance was compared with the conventional CCA-based method. The presented results indicated that the CNN is robust in decoding SSVEP across different ISDs and outperforms CCA in all cases. The average accuracy increased by over 10% using CNN on the closest ISD, which is the most challenging case with the most significant completing stimuli.

The evidence suggests that the CNN can be trained independent of the ISD resulting in a model that can generalize to ISDs that are not seen in its training data. Moreover, for real-world application scenarios, it would be more practical and feasible to implement a training scheme with the stimuli configuration as in Case 1 investigated here, wherein the model is trained with data obtained from interfaces with small ISDs, and then applied or run on interfaces with larger ISDs. These results are especially beneficial for practical applications developed on virtual reality or augmented reality platforms where the stimuli would be very closely spaced in initial training and calibration. From an interface design perspective, this method provides more flexibility for application development as newly configured stimulus distances can be easily modified with a simple software update and retain the same CNN weights for inference.

The analysis on the computational load shows that the CNN can be trained in a very short time (6s) and the execution time is also sufficiently short (mean inference time of 1.3ms on a mainstream PC). The proposed experimental protocol for collecting the training data was approximately 12 minutes. For practical user-customized systems, a shorter training time is desirable and the proposed experimental protocol serves this purpose. A 3-channel setup with O1, O2 and Oz shows the ease of setup of the proposed system. To accommodate higher number of channels, and more SSVEP classes the proposed CNN model can be easily modified at the input and output layers respectively. This will be studied in a future study. The proposed model and training methodology provides a simple and effective system for detecting SSVEP.

Acknowledgment

We thank Dr. Ali Ghodsi for his valuable suggestions. We also thank the participants who took part in this study.

References

[1] Danhua Zhu, Jordi Bieger, Gary Garcia Molina, and Ronald M. Aarts. A survey of stimulation methods used in SSVEP-based BCIs, 2010.

[2] Kian B. Ng, Andrew P. Bradley, and Ross Cunnington. Stimulus specificity of a steady-state visual-evoked potential-based brain-computer interface. *Journal of Neural Engineering*, 2012.

[3] Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U Rajendra Acharya. Deep learning for healthcare applications based on physiological signals: a review. *Computer Methods and Programs in Biomedicine*, 2018.

[4] No Sang Kwak, Klaus Robert Müller, and Seong Whan Lee. A convolutional neural network for steady state visual evoked potential classification under ambulatory environment. *PLoS ONE*, 2017.

[5] Nicholas Waytowich, Vernon J. Lawhern, Javier O. Garcia, Jennifer Cummings, Josef Faller, Paul Sajda, and Jean M. Vettel. Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials. *Journal of neural engineering*, 2018.

[6] Y. Wang, Y.-T. Wang, and T.-P. Jung. Visual stimulus design for high-rate SSVEP BCI. *Electronics Letters*, 46(15):1057, 2010.

[7] Masaki Nakanishi, Yijun Wang, Yu Te Wang, Yasue Mitsukura, and Tzyy Ping Jung. Generating visual flickers for eliciting robust steady-state visual evoked potentials at flexible frequencies using monitor refresh rate. *PLoS ONE*, 2014.

[8] Anna Duszyk, Maria Bierzyńska, Zofia Radzikowska, Piotr Milanowski, Rafal Kus, Piotr Suffczyński, Magdalena Michalska, Maciej Labęcki, Piotr Zwoliński, and Piotr Durka. Towards an optimization of stimulus parameters for brain-computer interfaces based on steady state visual evoked potentials. *PLoS ONE*, 9(11):1–11, 2014.

[9] Vernon J Lawhern, Amelia J Solon, and Nicholas R Waytowich. EEGNet : a compact convolutional neural network for EEG-based brain computer interfaces. 2018.

[10] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015.

[11] Zhonglin Lin, Changshui Zhang, Wei Wu, and Xiaorong Gao. Frequency recognition based on canonical correlation analysis for SSVEP-Based BCIs. *IEEE Transactions on Biomedical Engineering*, 54(6):1172–1176, 2007.